Predictive Determination of Car Prices Dependent upon Factors of Importance

Ishan Agrahar, Prajwal Bhandari, Noah Tobias

Background and Introduction

The act of buying a car is a major decision and used car dealerships aim to create an environment that turns potential buyers into actual ones. Buyers often scrutinize car features more closely in used markets, so maximizing profit depends on understanding which features drive purchase decisions.

This report explores two key research questions under the broader goal of building a model to make inferences and predictions about sale prices:

1. Are any of the features measured useful in predicting the sale prices of used cars? If so, what are they and how do they affect the sale price of used cars?
2. How can the measured features be used to predict the sale price of the used cars?

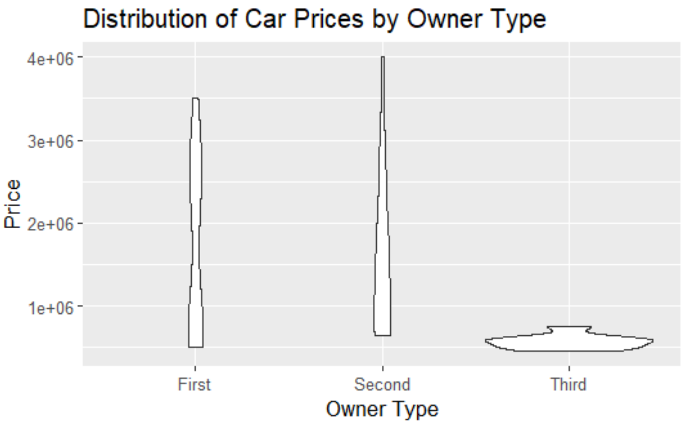
We analyze a dataset of 90 cars from an Asian dealership with ten distinct variables each (Appendix A):

- Unique Car Identifier (carID), Kilometers Driven (KmDriven), Gas or Diesel (FuelType), Manual or Automatic (Transmission), Fuel Efficiency (Mileage), Engine Capacity (Engine), Maximum Power output (Power), Number of Seats (Seats), Age of Vehicle (Age), Sale Price of Vehicle (Price).

To analyze the data we will use an 80-20 train-test split on the ninety observations and apply standardized statistical methods endorsed by Professor Bulathsinhala, ensuring reproducibility of results using the same approach and dataset.

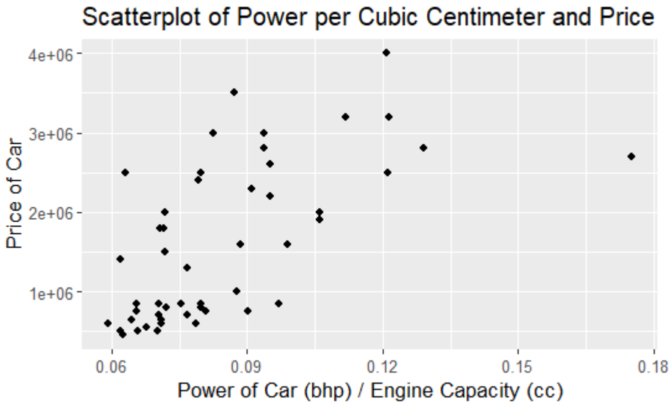
Exploratory Data Analysis

Before modeling, we first explore the data through univariate and multivariate analysis. One key aspect is the distribution of car prices by owner type, as the number of previous owners can influence perceived car value and provide useful insights.



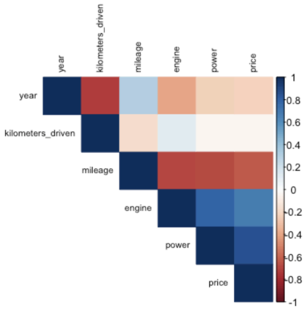
The violin plot shows the distribution of the prices based on owner type. The first and second ownership types have pretty uniform shape, while the third type is much more compact. It's harder to tell if there is any significant difference between the first owner and second owner types, but it's clear that the third owner type sells for much lower than the other two groups.

Buyers may also want to consider the amount of power a car can put out for a unit of engine capacity. The scatter plot to the left reveals what may be seen as an obvious relationship between the two variables. Generally the prices are increasing as the power/cc increases with one high leverage point at 0.18 bhp/cc. What is also interesting is that the variability of the prices for the cars seems to be increasing as a function of bhp/cc as well. It wouldn't be the best decision to fit a linear model in this case, yet it still highlights insights we can gain from the data, even if very qualitative.



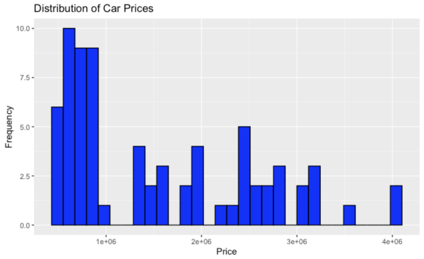
Statistical Methods and Analysis

Correlation Plot

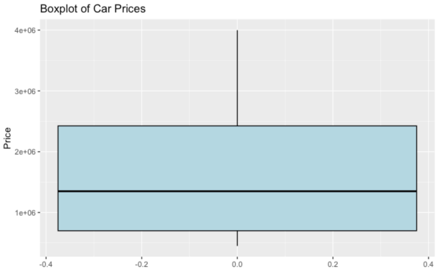


The correlation matrix shows a strong positive relationship between engine size and power, and a strong negative correlation between year and kilometers driven—newer cars typically have fewer kilometers. Mileage is negatively correlated with price, engine, and power, suggesting higher-mileage cars are generally cheaper. These high correlations among independent variables may indicate multicollinearity, which will be addressed during modeling.

Distribution of Car Prices



The histogram of car prices reveals a right-skewed distribution, where a majority of cars fall within the lower price range, with fewer high-priced vehicles. This suggests that applying a log transformation to price may help normalize the distribution, making it more suitable for predictive modeling.



The box plot shows a wide spread in car prices, with a high upper range likely due to luxury models. These may skew predictions, so further analysis is needed to decide whether to remove, adjust, or model them separately.

Feature Engineering Discussion:

To enhance model performance, we engineered new features:

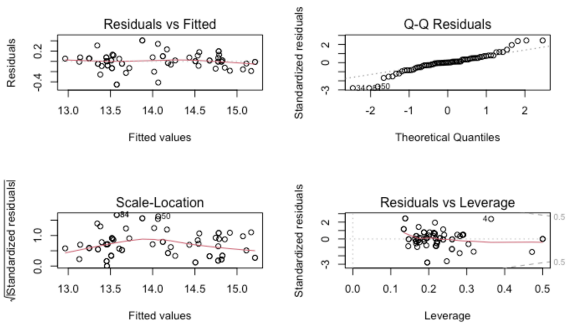
- **Quadratic Age** captures diminishing depreciation effects.
- **Log Price** normalizes skewed price data for better linearity.
- **Log Kilometers Driven** reduces skew from high-mileage cars.
- **Power per CC** reflects the efficiency of engine output, aiding price prediction.

Model Selection Discussion

Using stepwise selection based on AIC, we identified a subset of predictors that best explain car prices while avoiding overfitting (Process: Appendix C). The final model call (in R) includes the following variables (Full Summary Appendix B):

```
lm(log_price ~ log_kmdriven + fuel_type + transmission + mileage + power + brand)
```

Model Assumption Checks Discussion



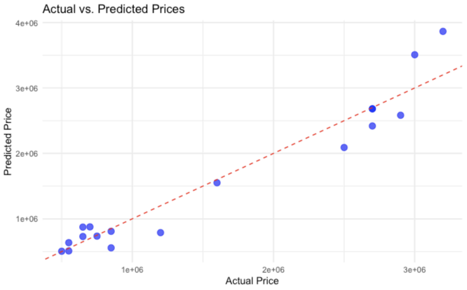
Normality of Residuals: The Q-Q plot shows residuals are roughly normal, with slight tail deviations suggesting minor issues at extreme values. Robust methods or transformations may help.

Homoscedasticity: Residuals show mostly constant variance, though the Scale-Location plot hints at mild instability for some fitted values.

Independence: If the cars weren't randomly sampled, generalizability may be limited. Independence holds if each price is unaffected by others.

Influential Points: A few high-leverage points (e.g., 50, 84, 63) don't exceed Cook's distance but merit review.

Multicollinearity: All VIFs are under 5, indicating no significant multicollinearity.



Prediction and Model Validation Discussion

The Actual vs. Predicted plot shows a strong linear fit, indicating the model captures pricing trends well across ranges, with only minor deviations at higher prices.

Performance Metrics:

- **RMSE:** ₹279,245.8, **MAE:** ₹202,082.1, **R²:** 0.9249

Conclusions

The first research question is which features are useful in predicting the sale prices of used cars and how do they affect the sale price of used cars? Referencing the summary in Appendix B, variables that were chosen by bi-directional AIC stepwise regression were Kilometers Driven, Fuel Type, Transmission Type, Mileage, Power, and Brand. The following relationships in predicting sales price are based on the sign of the coefficients in the stepwise regression model (Appendix B). The variables that have a negative correlation to sale price are petrol fuel type, manual transmission type, mileage, and certain car brands. This indicates that as the value of these variables increase, the sale price of the used vehicle will decrease. Examples of this is if the fuel type is petrol rather than diesel, the sale price is expected to decrease; another example is if the mileage increases, the sale price is expected to decrease. To specify the car brands that have a negative correlation to sale price, the brands are Ford, Hyundai, Mahindra, Maruti, Mercedes, Tata, Toyota, and Volkswagen. Note that these negative correlations are compared to a car that is not specified in its brand. The variables that have a positive correlation are kilometers driven, power and brand being BMW. An example application of this is if power increases, the sale price is expected to increase; if the brand is BMW the sale price is expected to increase.

The second research question is how can the features be used to predict the sale price of the used cars? For a prediction, substitute values associated with the car into this model. Note variables with I are binary variables: input a one in the position if the variable is satisfied, zero otherwise. (Summary: Appendix B)

$$\widehat{Sale\ Price} = e^{13.6369625} \times (kmdriven + 1)^{1.12799881213} \times e^{-0.2294106(I:Petrol)} \times e^{-0.3239128(I:Manual)}$$
$$\times e^{-0.0380747(mileage)} \times e^{0.0029584(power)} \times e^{0.0684343(I:BMW)} \times e^{-0.4725770(I:Ford)} \times e^{-0.4330813(I:Honda)}$$
$$\times e^{-0.7843374(I:Hyundai)} \times e^{-0.6573739(I:Mahindra)} \times e^{-0.5011400(I:Maruti)} \times e^{-0.0327345(I:Mercedes)}$$
$$\times e^{-0.5058603(I:Tata)} \times e^{-0.4122362(I:Toyota)} \times e^{-0.4865364(I:Volkswagen)}$$

Discussion (Considerations, limitations, etc)

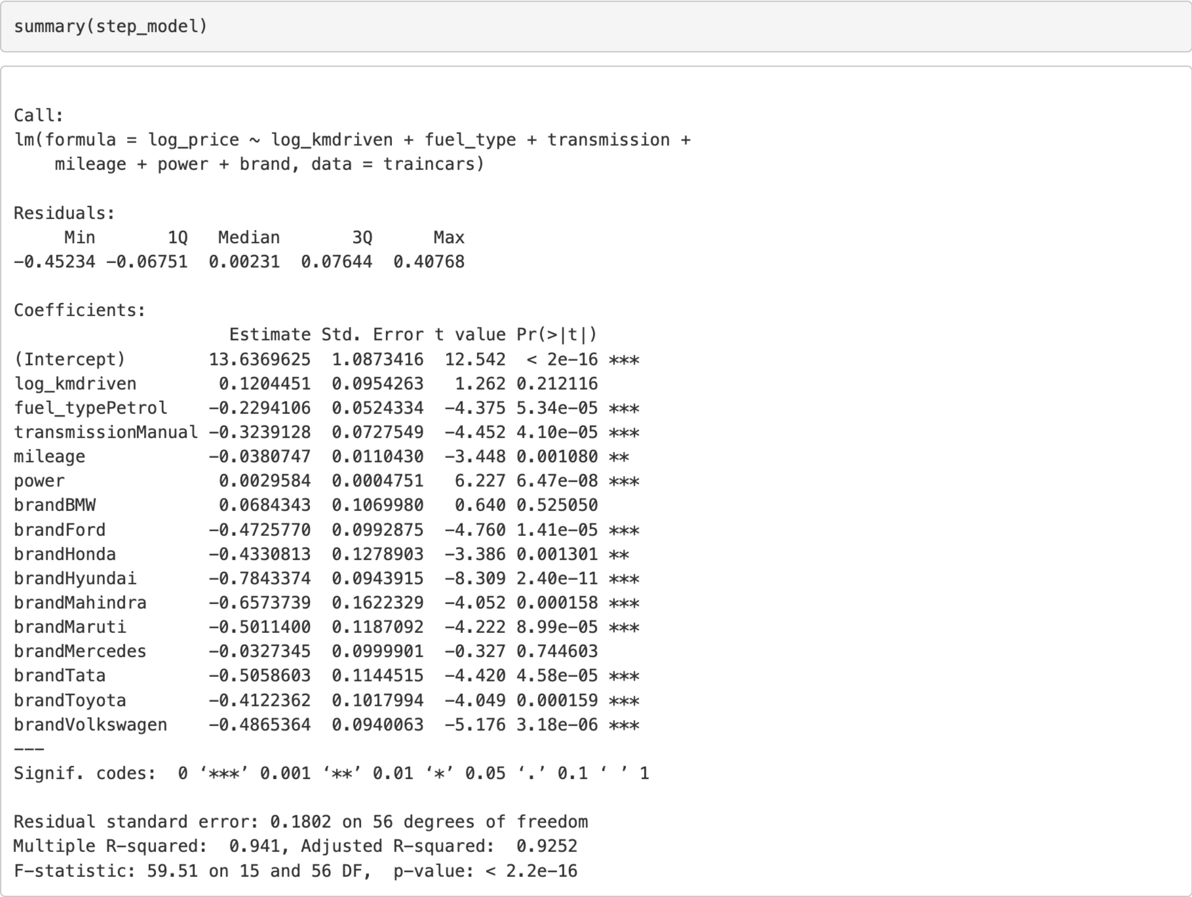
The biggest limitation of this report is that there is a lack of data. A total of ninety observations is not ideal under the guidelines that we are training our statistical model on seventy two observations. Throughout our analysis we had to check the validity assumptions for a linear model (linearity, normal, constant error variance). If this data is larger it is unclear whether these conditions would still be valid. Another limitation is that this model is trained on data from a dealership in an Asian country and therefore should not be generalized to other dealerships or other countries. On further considerations we are operating under the assumption that the data is independent; in the case this data was not independently sampled, the results of this report are at severe risk. Note there is a possibility that there are comparably good models to predict the sale price of used cars in the target population, however, through stepwise regression we are able to choose a final model which is the goal of the research questions. The future direction for this report is further research. This includes expanding the dataset with observations, diversifying available features, and further testing the effectiveness of this model or possibly creating other statistical models.

Appendix

Appendix A: The dataset contains ten distinct variables for each of the ninety car observations

- *carID* - A unique identification number of the car
- *KmDriven* - The total kilometers driven by the car
- *FuelType* - The type of fuel used by the car (Petrol or Diesel)
- *Transmission* - The transmission type of the car (Manual or Automatic)
- *Mileage* - The fuel efficiency of the car in kilometers per liter
- *Engine* - The engine capacity of the car in Cubic Centimeters.
- *Power* - The maximum power output of the car in bhp
- *Seats* - The number of seats available in the car
- *Age* - The age of the car (this year - manufactured year)
- *Price* - The selling price of the car in rupees

Appendix B: Full summary of the final linear regression model



```
summary(step_model)
```

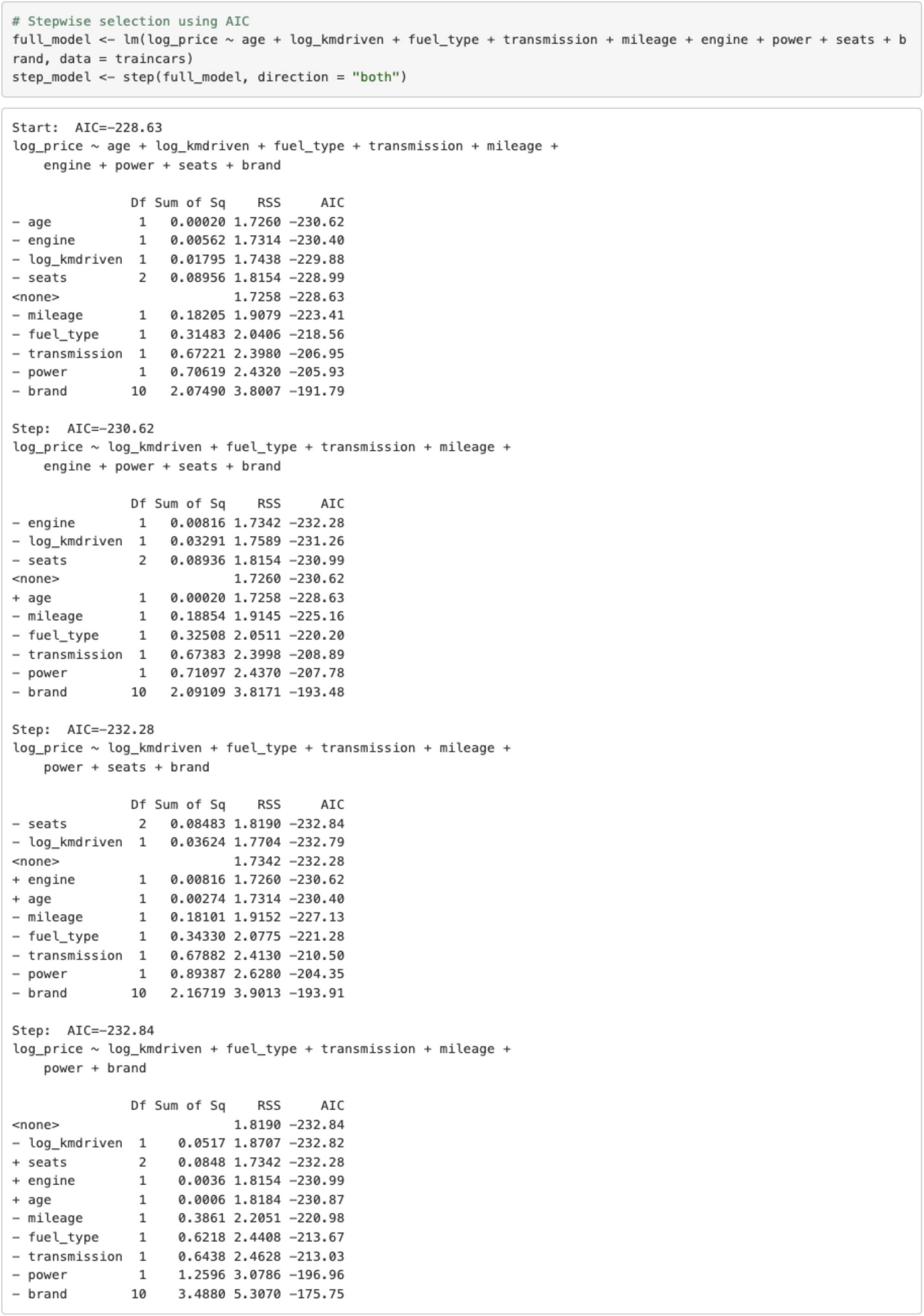
```
Call:
lm(formula = log_price ~ log_kmdriven + fuel_type + transmission +
    mileage + power + brand, data = traincars)

Residuals:
     Min       1Q   Median       3Q      Max
-0.45234 -0.06751  0.00231  0.07644  0.40768

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        13.6369625  1.0873416  12.542  < 2e-16 ***
log_kmdriven        0.1204451  0.0954263   1.262 0.212116
fuel_typePetrol    -0.2294106  0.0524334  -4.375 5.34e-05 ***
transmissionManual -0.3239128  0.0727549  -4.452 4.10e-05 ***
mileage            -0.0380747  0.0110430  -3.448 0.001080 **
power               0.0029584  0.0004751   6.227 6.47e-08 ***
brandBMW            0.0684343  0.1069980   0.640 0.525050
brandFord          -0.4725770  0.0992875  -4.760 1.41e-05 ***
brandHonda         -0.4330813  0.1278903  -3.386 0.001301 **
brandHyundai       -0.7843374  0.0943915  -8.309 2.40e-11 ***
brandMahindra      -0.6573739  0.1622329  -4.052 0.000158 ***
brandMaruti        -0.5011400  0.1187092  -4.222 8.99e-05 ***
brandMercedes      -0.0327345  0.0999901  -0.327 0.744603
brandTata          -0.5058603  0.1144515  -4.420 4.58e-05 ***
brandToyota        -0.4122362  0.1017994  -4.049 0.000159 ***
brandVolkswagen    -0.4865364  0.0940063  -5.176 3.18e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1802 on 56 degrees of freedom
Multiple R-squared:  0.941, Adjusted R-squared:  0.9252
F-statistic: 59.51 on 15 and 56 DF,  p-value: < 2.2e-16
```

Appendix C: Stepwise Selection Process for final model.



```
# Stepwise selection using AIC
full_model <- lm(log_price ~ age + log_kmdriven + fuel_type + transmission + mileage + engine + power + seats + b
rand, data = traincars)
step_model <- step(full_model, direction = "both")
```

```
Start:  AIC=-228.63
log_price ~ age + log_kmdriven + fuel_type + transmission + mileage + 
    engine + power + seats + brand

               Df Sum of Sq    RSS     AIC
- age           1   0.00020 1.7260 -230.62
- engine        1   0.00562 1.7314 -230.40
- log_kmdriven  1   0.01795 1.7438 -229.88
- seats         2   0.08956 1.8154 -228.99
<none>                      1.7258 -228.63
- mileage       1   0.18205 1.9079 -223.41
- fuel_type     1   0.31483 2.0406 -218.56
- transmission  1   0.67221 2.3980 -206.95
- power         1   0.70619 2.4320 -205.93
- brand        10   2.07490 3.8007 -191.79

Step:  AIC=-230.62
log_price ~ log_kmdriven + fuel_type + transmission + mileage + 
    engine + power + seats + brand

               Df Sum of Sq    RSS     AIC
- engine        1   0.00816 1.7342 -232.28
- log_kmdriven  1   0.03291 1.7589 -231.26
- seats         2   0.08936 1.8154 -230.99
<none>                      1.7260 -230.62
+ age           1   0.00020 1.7258 -228.63
- mileage       1   0.18854 1.9145 -225.16
- fuel_type     1   0.32508 2.0511 -220.20
- transmission  1   0.67383 2.3998 -208.89
- power         1   0.71097 2.4370 -207.78
- brand        10   2.09109 3.8171 -193.48

Step:  AIC=-232.28
log_price ~ log_kmdriven + fuel_type + transmission + mileage + 
    power + seats + brand

               Df Sum of Sq    RSS     AIC
- seats         2   0.08483 1.8190 -232.84
- log_kmdriven  1   0.03624 1.7704 -232.79
<none>                      1.7342 -232.28
+ engine        1   0.00816 1.7260 -230.62
+ age           1   0.00274 1.7314 -230.40
- mileage       1   0.18101 1.9152 -227.13
- fuel_type     1   0.34330 2.0775 -221.28
- transmission  1   0.67882 2.4130 -210.50
- power         1   0.89387 2.6280 -204.35
- brand        10   2.16719 3.9013 -193.91

Step:  AIC=-232.84
log_price ~ log_kmdriven + fuel_type + transmission + mileage + 
    power + brand

               Df Sum of Sq    RSS     AIC
<none>                      1.8190 -232.84
- log_kmdriven  1    0.0517 1.8707 -232.82
+ seats         2    0.0848 1.7342 -232.28
+ engine        1    0.0036 1.8154 -230.99
+ age           1    0.0006 1.8184 -230.87
- mileage       1    0.3861 2.2051 -220.98
- fuel_type     1    0.6218 2.4408 -213.67
- transmission  1    0.6438 2.4628 -213.03
- power         1    1.2596 3.0786 -196.96
- brand        10    3.4880 5.3070 -175.75
```